Final Year Project

EVENT SPOTTING IN SOCCER GAMES USING A SCALABLE DATASET

Guide: Dr. S.Sudha

Team Members

Hari Krishna 2015103017

Sreevatsan Agoramurthy 2015103023

Sribashyam Sriranjan Srinivas 2015103049

Review 1

**PROBLEM STATEMENT**

Sports is a lucrative sector, with large amounts of money being invested on players and teams. After merchandising, TV broadcast rights are the second major revenue stream for a soccer club. Even though the main scope of soccer broadcast is entertainment, such videos are also used by professionals to generate statistics, analyze strategies, and scout new players. In order to get such statistics, professional analysts watch a lot of broadcasts and identify all the events that occur within a game.

Our objective is to develop a model that can perform localization of very-sparse events in long videos and annotate them accordingly. We achieve these annotations by recognizing three major classes of events, namely

* Card
* Substitution
* Goal

**INTRODUCTION**

Sports is a lucrative sector, with large amounts of money being invested on players and teams. The global sports market is estimated to generate an annual revenue of $91 billion, whereby the European soccer market contributes about $28.7 billion, from which $15.6 billion alone come from the Big Five European soccer leagues (EPL, La Liga, Ligue 1, Bundesliga and Serie A). After merchandising, TV broadcast rights are the second major revenue stream for a soccer club. Even though the main scope of soccer broadcast is entertainment, such videos are also used by professionals to generate statistics, analyze strategies, and scout new players. In order to get such statistics, professional analysts watch a lot of broadcasts and identify all the events that occur within a game. Automated methods for sports video understanding can help in the localization of the salient actions of a game. Several companies such as Reely are trying to build automated methods to understand sports broadcasts and would benefit from a large-scale annotated dataset for training and evaluation. Many recent methods exist to solve generic human activity localization in video focusing on sports. However, detecting soccer actions is a difficult task due to the sparsity of the events within a video. Soccer broadcast understanding can thus be seen as a sub-problem of video understanding, focusing on a vocabulary of sparse events defined within its own context.

Sports analytics are a collection of relevant, historical, statistics that when properly applied can provide a competitive advantage to a team or individual. Through the collection and analyzation of these data, sports analytics inform players, coaches and other staff in order to facilitate decision making both during and prior to sporting events. The term "sports analytics" was popularized in mainstream sports culture following the release of the 2011 film, Moneyball, in which Oakland Athletics General Manager Billy Beane (played by Brad Pitt) relies heavily on the use of analytics to build a competitive team on a minimal budget. There are two key aspects of sports analytics - on-field and off-field analytics. On-field analytics deals with improving the on-field performance of teams and players. It digs deep into aspects such as game tactics and player fitness. Off-field analytics deals with the business side of sports. Off-field analytics focuses on helping a sport organization or body surface patterns and insights through data that would help increase ticket and merchandise sales, improve fan engagement, etc. Off-field analytics essentially uses data to help rightsholders take better decisions that would lead to higher growth and increased profitability. In this paper, we are focusing on the on-field analytics. More specifically , we are trying to develop a model that will provide us with data that could be used later for on-field analytics.

Coming to the topic of training a model for our proposed work, we need to identify the right kind of architecture to suit our requirements. In a traditional CNN architecture, according to the universal approximation theorem, given enough capacity, we know that a feedforward network with a single layer is sufficient to represent any function. However, the layer might be massive and the network is prone to overfitting the data. Therefore, there is a common trend in the research community that our network architecture needs to go deeper. However, increasing network depth does not work by simply stacking layers together. Deep networks are hard to train because of the notorious vanishing gradient problem — as the gradient is back-propagated to earlier layers, repeated multiplication may make the gradient infinitively small. As a result, as the network goes deeper, its performance gets saturated or even starts degrading rapidly. The core idea of ResNet is introducing a so-called “identity shortcut connection” that skips one or more layers. This is our main reason to use ResNet for our requirements.

Later, we use a shallow pooling neural network to work with the reduced set of features, ones that we had obtained from the ResNet model. We use a convolutional neural network (CNN) to achieve this . The main component of this architecture, NetVLAD, is a new generalized VLAD layer, inspired by the "Vector of Locally Aggregated Descriptors" image representation commonly used in image retrieval. The layer is readily pluggable into any CNN architecture and amenable to training via backpropagation. Using these various processed we attempt to annotate the videos by spotting events and dividing them into three main classes.

**LITERATURE SURVEY**

* **Sports Analytics**

Many automated sports analytics methods have been developed in the computer vision community to understand sports broadcasts. They produce statistics of events within a game by either analyzing camera shots or semantic information. Ekin et al. present a cornerstone work for game summarization based on camera shot segmentation and classification, followed by Ren et al. who focus on identifying video production patterns. Huang et al. analyze semantic information to automatically detect goals, penalties, corner kicks, and card events. Tavassolipour et al. use Bayesian networks to summarize games by means of semantic analysis. More recent work in this category focuses on deep learning pipelines to localize salient actions in soccer videos. Baccouche et al. use a Bag-of-Words (BOW) approach with SIFT features to extract visual content within a frame. They use such representations to train a Long Short Term Memory (LSTM) network that temporally traverses the video to detect the main actions. Jiang et al. pro- pose a similar methodology using Convolution Neural Networks (CNN) to extract global video features rather than local descriptors. They also use a play-break structure to generate candidate actions. Tsagkatakis et al. present a two-stream approach to detect goals, while Homayounfar et al. recently resent a deep method for sports field localization, which is crucial for video registration purposes. The main impediment for all these works is the lack of reference datasets/benchmarks that can be used to evaluate their performance at large-scale and standardize their comparison. They all use small and custom-made datasets, which contain a few dozen soccer games at most. We argue that intelligent sports analytics solutions need to be scalable to the size of the problem at hand. Therefore, to serve and support the development of such scalable solutions, we propose a very large soccer-centric dataset that can be easily expanded and enriched with various types of annotations.

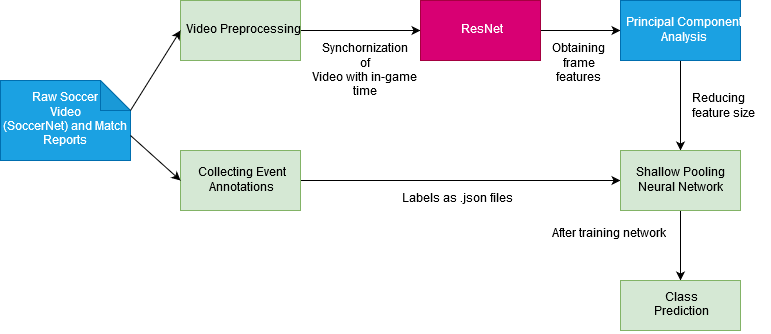
* **Activity Recognition**

Activity recognition focuses on understanding videos by either detecting activities or classifying segments of video according to a predefined set of human-centric action classes. A common pipeline consists of proposing temporal segments, which are in turn further pruned and classified. Common methods for activity classification and detection make use of dense trajectories, action’s estimation, Recurrent Neural Networks (RNN), tubelets, and handcrafted features. In order to recognize or detect activities within a video, a common practice consists of aggregating local features and pooling them, looking for a consensus of characteristics. While naive approaches use mean or maximum pooling, more elaborate techniques such as Bag-of-Words (BOW), Fisher Vector (FV), and VLAD look for a structure in a set of features by clustering and learning to pool them in a manner that improves discrimination. Recent works extend those pooling techniques by incorporating them into Deep Neural Network (DNN) architectures, namely NetFV, SoftDBOW, and NetVLAD. By looking for correlations between a set of primitive action representations, ActionVLAD has shown state- of-the-art performance in several activity recognition benchmarks. To further improve activity recognition, recent works focused on exploiting context, which represent and harness information in both temporal and/or spatial neighborhood, or on attention, which learns an adaptive conﬁdence score to leverage this surrounding information. In this realm, Caba Heilbron et al. develop a semantic context encoder that exploits evidence of objects and scenes within video segments to improve activity detection effectiveness and efﬁciency. Miech et al. , winners of the ﬁrst annual Youtube 8M challenge, show how learnable pooling can produce state-of-the-art recognition performance on a very large benchmark, when recognition is coupled with context gating. More recently, several works use temporal context to localize activities in videos or to generate proposals. Furthermore, Nguyen et al. present a pooling method that uses spatio-temporal attention for enhanced action recognition, while Pei et al. use temporal attention to gate neighboring observations in a RNN framework. Note that attention is also widely used in video captioning. Activity recognition and detection methods are able to provide good results for these complicated tasks. However, those methods are based on DNNs and require large-scale and rich datasets to learn a model. By proposing a large scale dataset focusing on event spotting and soccer, we encourage algorithmic development in those directions.

* **Datasets**

Multiple datasets are available for video understanding, especially for video classiﬁcation. They include Hollywood2 and HMDB, both focusing on movies; MPII Cooking, focusing on cooking activities; UCF101, for classiﬁcation in the wild; UCF Sports, Olympics Sports and Sports-1M, focusing on sports; Youtube-8M and Kinetics, both tackling large-scale video classiﬁcation in the wild. They are widely used in the community but serve the objective of video classiﬁcation rather than activity localization. The number of benchmark datasets focusing on action localization is much smaller. THUMOS14 is the ﬁrst reasonably scaled benchmark for the localization task with a dataset of 413 untrimmed videos, totaling 24 hours and 6,363 activities, split into 20 classes. MultiTHUMOS is a subset of THUMOS, densely annotated for 65 classes over unconstrained internet videos. ActivityNet tackles the issue of general video understanding using a semantic ontology, proposing challenges in trimmed and untrimmed video classiﬁcation, activity localization, activity proposals and video captioning. ActivityNet 1.3 provides a dataset of 648 hours of untrimmed videos with 30,791 activity candidates split among 200 classes. It is so far the largest localization benchmark in terms of total duration. Charades is a recently compiled benchmark for temporal activity segmentation that crowd-sources the video capturing process. After collecting a core set of videos from YouTube, they use AMT to augment their data by recording them at home. This dataset consists of a total of 9,848 videos and 66,500 activities. More recently, Google proposed AVA as a dataset to tackle dense activity understanding. They provide 57,600 clips of 3 seconds duration taken from featured ﬁlms, annotated with 210,000 dense spatio-temporal labels across 100 classes, for a total of 48 hours of video. While the main task of AVA is to classify these 3 seconds segments, such dense annotation can also be used for detection. Within the multimedia community, TRECVID has been the reference benchmark for over a decade. They host a “Multimedia Event Detection” (MED) and a “Surveillance Event Detection” (SED) task every year, using the HAVIC dataset. These tasks focus on ﬁnding all clips in a video collection that contain a given event, with a textual deﬁnition, in multimedia and surveillance settings. Also, Ye et al. propose EventNet, a dataset for event retrieval based on a hierarchical ontology, similar to ActivityNet. We argue that these two datasets both focus on large-scale information retrieval rather than video understanding. Hence, we us SoccerNet ,a scalable and soccer-focused dataset for event spotting. It contains 500 games with almost 764 hours of video which makes it one of the largest dataset in term of total duration and number of instances per class. With an average of one event every 6.9 minutes, our dataset has a sparse distribution of events in long untrimmed videos, which makes the task of localization more difﬁcult. The annotations are obtained within one minute at no cost by parsing sports websites, and further reﬁned in house to one second resolution.

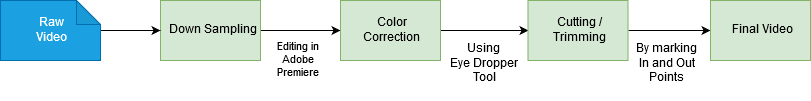
**SYSTEM ARCHITECTURE**



*Main architecture*

**MODULE DESCRIPTION**

1. Video Preprocessing



**Input:** Raw video footage of soccer games

**Output:** Two 45 min halves of the soccer game in drive

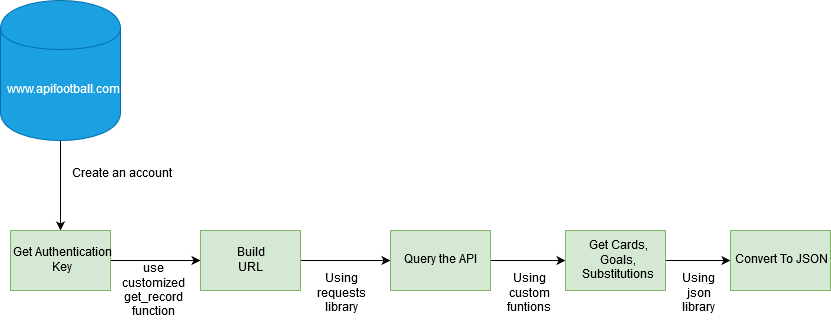
**Psuedocode:**

* + - Downsample the video to 224x224 resolution
    - Using Adobe Premium Pro, remove the media coverage and replays present in the footage; the timestamp in the video should match with the actual in-game time
    - Color correct the videos as different soccer leagues have different color schemes and we need uniformity through all the videos
    - Lock the frame rate of every video to 25 frames/second

**Description:**

Each game is composed of 2 untrimmed videos, one for each half period. The videos come from online providers, in a variety of encodings (MPEG, H264), containers (MKV, MP4, and TS), frame rates (25 to 50 fps), and resolutions (SD to FullHD). To standardize these videos, the videos are trimmed at the game start, resized and cropped at a 224 × 224 resolution, and uniﬁed at 25fps. The dataset has videos for a total duration of 45 hours.

2. Collecting Event Annotations



**Input:** Python script with API to crawl the web

**Output:** .JSON files containing event time stamps

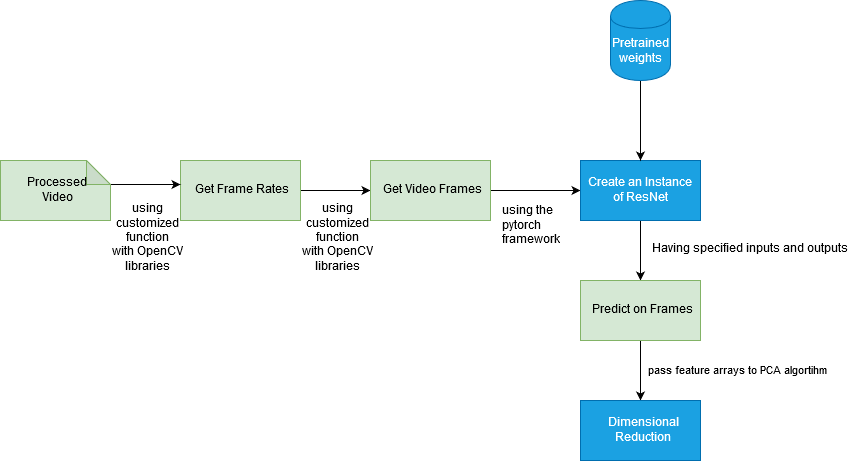
**Psuedocode:**

* + - Register with [www.apifootball.com](http://www.apifootball.com/).; get authentication key
    - API endpoint has following parameters: League ID, Date\_from & Date\_to, Team\_Name
    - Build the URL using get-record() function (send parameters here)
    - Query the API to collect the details of the soccer game
    - Define three separate functions- getcards(), getgoals(), getsubstitutions() in order to seperate only the relevant data

**Description:**

We use a python script to crawl the web in order to collect match reports. These match reports are to be later used to annotate the events in the video. An API helps us to achieve this by taking in the parameters – League, Date, Team Name and returning us the match report. Next, our python script classifies the events that have been mentioned in the match report into – cards, goals & substitutions.

3. Feeding videos into ResNet



**Input:** 45min video

**Output:**  Frame features for the video

**Psuedocode:**

* We use getFPS() in order to get the FPS of the video
* GetVideoFrames() to get the video frame arguments: VideoPath, GameStartIndex, GameEndIndex, Height, Width using the handlers such as OpenCV or SciKit-Learn
* The output will be 4D Numpy array per second of a video reading the metadata
* Create an instance of ResNet with input as dimensions of video frames and output as the fc1000 layer in ResNet
* We choose every 25th frame basically meaning the frame after every second
* Using resnet, predict values on the set of the frames
* Return feature arrays

4. Reduce dimensionality using PCA

**Input:** 5.5M feature array

**Output:** Feature array with reduced dimensionality

**Description :**

Since there are 2048 feature per single frame the number of total frame features are 5.5 million features (2048 frame features \* 60 seconds \* 45mins).

Therefore, we use PCA to reduce the dimension to 512 features

5. Shallow Pooling Neural Network



**Input:** Frame features of the video

**Output:** A model that we can use to predict the type of event

**Pseudocode :**

* Find features which correspond to event times extracted from annotated Match Reports
* Feed it through a simple Neural Network with very few layers
* Use NetVLAD pooling technique
* Sigmoid Activation Layer, Adam optimizer minimizing binary cross-entropy loss function
* Evaluate output using mAP(mean average Precision)

**IMPLEMENTATION DETAILS – 30%**

Collecting Soccer Video Feed

The soccer videos are collected from the scalable dataset. The videos are of 2 untrimmed videos. We collect these videos and trim these videos at the start of the game time and the end of the game time, thus giving us a trimmed video of just the game (no commercials and post-match talk show). Next, we standardize the videos by cropping them all to a low resolution of 224 x 224 pixels and we lock the frame rate to 25 fps (frames per second). This makes it much easier for us to process when we pass the videos to the ResNet Model.



*Screenshot of a match after being standardized*

Collecting match reports

We use a python script to crawl the web and obtain match reports. We achieve this by using an API from [www.apifootball.com](http://www.apifootball.com/).

****

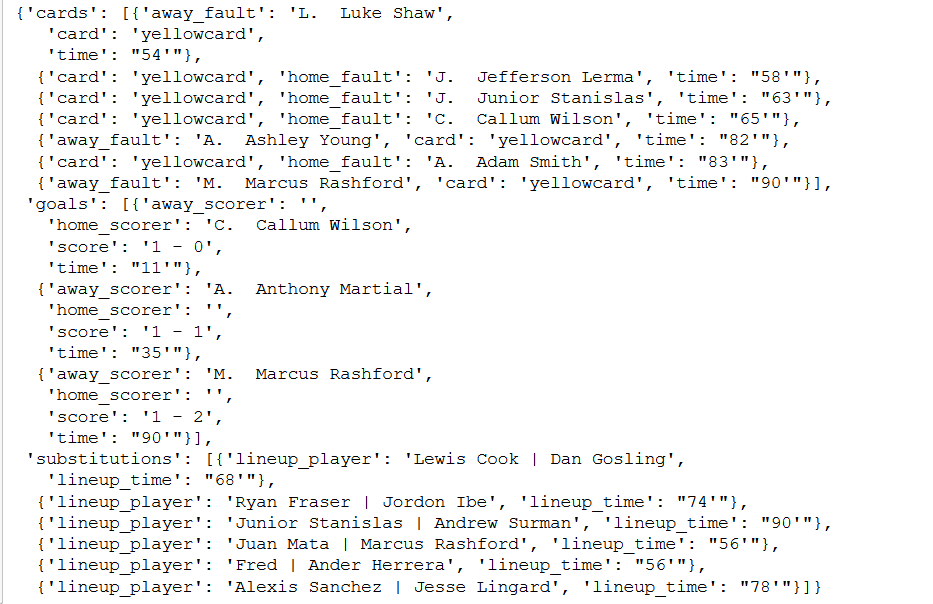
*Screenshot of our algorithm using APIFootball*

We pass in the parameters - League, Date, Team Name and we then obtain the corresponding match reports.

****

*Screenshot of functions for getting the relevant details from the crawled data*

We then obtain the records of the event with the corresponding timestamp. This is what will help us annotate the videos, the ones that are preprocessed, and then later use these videos to develop a model.

****

*Screenshot of the records obtained, each event mentioned has a timestamp along with it*

**PERFORMANCE METRICS**

For evaluation of our model we calculate the Recall, Precision and mean Average Precision for each class of event identified and also a mean Average Precision over all the classes.

Recall = True Positives/ (True Positives+ False Negatives)

Precision = True Positives/ (True Positives+ False Positives)

mean Average Precision = Sum of all precisions/No.of Precisions

**DATASET**

We use the SoccerNet Dataset consisting of video footage from soccer games across the top five leagues in Europe along with European Cups. The dataset used comes in two variations, one of about 119 GB, of lower quality. The other of Higher quality, of > 2TB in size. The accompanying labels of the dataset. Dataset would be video feeds of games, annotated by the labels.

**REFERENCES**

* P. Oskouie, S. Alipour, and A.-M. Eftekhari-Moghadam, “Multimodal feature extraction and fusion for semantic mining of soccer video: A survey,” *Artif. Intell. Rev.*, vol. 42, no. 2, pp. 173–210, 2015.
* Q. Ye, Q. Huang, S. Jiang, Y. Liu, and W. Gao, “Jersey number detection in sports video for athlete identification,” *Proc. SPIE*, vol. 5960, pp. 1599–1606, May 2015.
* M1. Boyar, Ö. Alan, S. Akpinar, O. Sabuncu, N. K. Çiçekli, and F N. Alpaslan, “Event boundary detection using audio-visual features and Web-casting texts with imprecise time information,” in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2014, pp. 578–583.
* C. Xu, J. Wang, H. Lu, and Y. Zhang, “A novel framework for semantic annotation and personalized retrieval of sports video,” *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 421–436, Apr. 2014.
* C. Xu, J. Wang, K. Wan, Y. Li, and L. Duan, “Live sports event detection based on broadcast video and Web-casting text,” in *Proc. 14th Annu. ACM Int. Conf. Multimedia*, 2015, pp. 221–230.
* H. Xu, T.-H. Fong, and T.-S. Chua, “Fusion of multiple asynchronous information sources for event detection in soccer video,” in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2005, pp. 1242–1245.
* Y. Gong, L. T. Sin, C. H. Chuan, H. Zhang, and M. Sakauchi, “Automatic parsing of TV soccer programs,” in *Proc. IEEE Int. Conf. Multimedia Comput. Syst.*, May 1995, pp. 167–174.
* X. Qian, H. Wang, G. Liu, and X. Hou, “HMM based soccer video event detection using enhanced mid-level semantic,” Multimedia Tools Appl., vol. 60, no. 1, pp. 233–255, 2012.
* H. Pan, P. van Beek, and M. I. Sezan, “Detection of slow-motion replay segments in sports video for highlights generation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 3. May 2001, pp. 1649–1652.
* Y. Yang, S. Lin, Y. Zhang, and S. Tang, “Highlights extraction in soccer videos based on goal-mouth detection,” in *Proc. 9th IEEE Int. Symp. Signal Process. Appl.*, Feb. 2007, pp. 1–4.
* Y. Li, G. Liu, and X. Qian, “Ball and field line detection for placed kick refinement,” in *Proc. Goal Congr. Intell. Syst.*, May 2009, pp. 404–407.